(54) Title: IDENTIFICATION OF NUCLEOTIDES, AMINO ACIDS, OR CARBOHYDRATES BY MASS SPECTROMETRY

(57) Abstract

A method for correlating a peptide fragment mass spectrum with amino acid sequences derived from a database is provided. A peptide is analyzed by a tandem mass spectrometer to yield a peptide fragment mass spectrum. A protein sequence database or a nucleotide sequence database is used to predict one or more fragment spectra for comparison with the experimentally-derived fragment spectrum. In one embodiment, sub-sequences of the sequences found on the database which define a peptide having a mass substantially equal to the mass of the peptide analyzed by the tandem mass spectrometer are identified as candidate sequences. For each candidate sequence, a plurality of fragments of the sequence are identified and the masses and m/z ratios of the fragments are predicted and used to form a predicted mass spectrum. The various predicted mass spectra are compared to the experimentally derived fragment spectrum using a closeness-of-fit measure, preferably calculated with a two-step process, including a calculation of a preliminary score end, for the highest-scoring predicted spectra, calculation of a correlation function.

IDENTIFICATION OF NUCLEOTIDES, AMINO ACIDS, OR CARBOHYDRATES BY MASS
SPECTROMETRY

Related Application

The present application is a continuation-in-part
of U.S. Serial No. 08/212,433, filed March 14, 1994, which is
incorporated herein by reference.

Background Of The Invention

A number of approaches have been used in the past
for applying the analytic power of mass spectrometry to
peptides. Tandem mass spectrometry (MS/MS) techniques have
been particularly useful. In tandem mass spectrometry, the
peptide or other input (commonly obtained from a
chromatography device) is applied to a first mass spectrometer
which serves to select, from a mixture of peptides, a target
peptide of a particular mass. The target peptide is then
activated or fragmented to produce a mixture of the "target"
or parent peptide and various component fragments, typically
peptides of smaller mass. This mixture is then transmitted to
a second mass spectrometer which records a fragment spectrum.
This fragment spectrum will typically be expressed in the form
of a bar graph having a plurality of peaks, each peak

2

indicating the mass-to-change ratio (m/z) of a detected
fragment and having an intensity value.

Although the bare fragment spectrum can be of some
interest, it is often desired to use the fragment spectrum to
identify the peptide (or the parent protein) which resulted in
the fragment mixture. Previous approaches have typically
involved using the fragment spectrum as a basis for
hypothesizing one or more candidate amino acid sequences.
This procedure has typically involved human analysis by a
skilled researcher, although at least one automated procedure
has been described. John Yates, III, et al., "Computer Aided
Interpretation of Low Energy MS/MS Mass Spectra of Peptides"
Techniques In Protein Chemistry II (1991), pp. 477-485,
incorporated herein by reference. The candidate sequences can
then be compared with known amino acid sequences of various
proteins in the protein sequence libraries.

The procedure which involves hypothesizing
candidate amino acid sequences based on fragment spectra is
useful in a number of contexts but also has certain
difficulties. Interpretation of the fragment spectra so as to
produce candidate amino acid sequences is time-consuming,
often inaccurate, highly technical and in general can be
performed only by a few laboratories with extensive experience
in tandem mass spectrometry. Reliance on human interpretation
often means that analysis is relatively slow and lacks strict
objectivity. Approaches based on peptide mass mapping are
limited to peptide masses derived from an intact homogenous
protein generated by specific and known proteolytic cleavage
and thus are not generally applicable to mixtures of proteins.

Accordingly, it would be useful to provide a system
for correlating fragment spectra with known protein sequences
while avoiding the delay and/or subjectivity involved in
hypothesizing or deducing candidate amino acid sequences from
the fragment spectra.

3

Summary Of The Invention

According to the present invention, known amino acid sequences, e.g., in a protein sequence library, are used to calculate or predict one or more candidate fragment spectra. The predicted fragment spectra are then compared with an experimentally-derived fragment spectrum to determine the best match or matches. Preferably, the parent peptide, from which the fragment spectrum was derived has a known mass. Sub- sequences of the various sequences in the protein sequence library are analyzed to identify those sub-sequences corresponding to a peptide whose mass is equal to (or within a given tolerance of) the mass of the parent peptide in the fragment spectrum. For each sub-sequence having the proper mass, a predicted fragment spectrum can be calculated, e.g., by calculating masses of various amino acid subsets of the candidate peptide. The result will be a plurality of candidate peptides, each with a predicted fragment spectrum. The predicted fragment spectra can then be compared with the fragment spectrum derived from the tandem mass spectrometer to identify one or more proteins having sub-sequences which are likely to be identical with the sequence of the peptide which resulted in the experimentally-derived fragment spectrum.

Brief Description Of The Drawings

Fig. 1 is a block diagram depicting previous methods for correlating tandem mass spectrometer data with sequences from a protein sequence library;

Fig. 2 is a block diagram showing a method for correlating tandem mass spectrometer data with sequences from a protein sequence library according to an embodiment of the present invention;

Fig. 3 is a flow chart showing steps for correlating tandem mass spectrometry data with amino acid sequences, according to an embodiment of the present invention;

4

Fig. 4 is a flow diagram showing details of a method for the step of identifying candidate sub-sequences of Fig. 3;

Fig. 5 is a fragment mass spectrum for a peptide of a type that can be used in connection with the present invention; and

Figs. 6A-6D are flow charts showing an analysis method, according to an embodiment of the present invention.

Description Of The Specific Embodiments

Before describing the embodiments of the present invention, it will be useful to describe, in greater detail, a previous method. As depicted in Fig. 1, the previous method is used for analysis of an unknown peptide 12. Typically the peptide will be output from a chromatography column which has been used to separate a partially fractionated protein. The protein can be fractionated by, for example, gel filtration chromatography and/or high performance liquid chromatography (HPLC). The sample 12 is introduced to a tandem mass spectrometer 14 through an ionization method such as electrospray ionization (ES). In the first mass spectrometer, a peptide ion is selected, so that a targeted component of a specific mass, is separated from the rest of the sample 14a. The targeted component is then activated or decomposed. In the case of a peptide, the result will be a mixture of the ionized parent peptide ("precursor ion") and component peptides of lower mass which are ionized to various states. A number of activation methods can be used including collisions with neutral gases (also referred to as collision induced dissolution). The parent peptide and its fragments are then provided to the second mass spectrometer 14c, which outputs an intensity and m/z for each of the plurality of fragments in the fragment mixture. This information can be output as a fragment mass spectrum 16. Fig. 5 provides an example of such a spectrum 16. In the spectrum 16 each fragment ion is represented as a bar graph whose abscissa value indicates the

mass-to-charge ratio (m/z) and whose ordinate value represents
intensity. According to previous methods, in order to
correlate a fragment spectrum with sequences from a protein
sequence library, a fragment sequence was converted into one
5 or more amino acid sequences judged to correspond to the
fragment spectrum. In one strategy, the weight of each of the
amino acids is subtracted from the molecular weight of the
parent ion to determine what might be the molecular weight of
a fragment assuming, respectively, each amino acid is in the
10 terminal position. It is determined if this fragment mass is
found in the actual measured fragment spectrum. Scores are
generated for each of the amino acids and the scores are
sorted to generate a list of partial sequences for the next
subtraction cycle. Cycles continue until subtraction of the
15 mass of an amino acid leaves a difference of less than 0.5 and
greater than -0.5. The result is one or more candidate amino
acid sequences 18. This procedure can be automated as
described, for example, in Yates III (1991) supra. One or
more of the highest-scoring candidate sequences can then be
20 compared 21 to sequences in a protein sequence library 20 to
try to identify a protein having a sub-sequence similar or
identical to the sequence believed to correspond to the
peptide which generated the fragment spectrum 16.

           Fig. 2 shows an overview of a process according to
25 the present invention. According to the process of Fig. 2, a
fragment spectrum 16 is obtained in a manner similar to that
described above for the fragment spectrum depicted in Fig. 1.
Specifically, the sample 12 is provided to a tandem mass
spectrometer 14. Procedures described below use a two-step
30 process to acquire ms/ms data. However the present invention
can also be used in connection with mass spectrometry
approaches currently being developed which incorporate
acquisition of ms/ms data with a single step. In one
embodiment ms/ms spectra would be acquired at each mass. The
35 first ms would separate the ions by mass-to-charge and the
second would record the ms/ms spectrum. The second stage of

6

ms/ms would acquire, e.g. 5 to 10 spectra at each mass
transformed by the first ms.

A number of mass spectrometers can be used
including a triple-quadruple mass spectrometer, a Fourier-
5   transform cyclotron resonance mass spectrometer, a tandem
time-of-flight mass spectrometer and a quadrupole ion trap
mass spectrometer. In the process of Fig. 2, however, it is
not necessary to use the fragment spectrum as a basis for
hypothesizing one or more amino acid sequences. In the
10   process of Fig. 2, sub-sequences contained in the protein
sequence library 20 are used as a basis for predicting a
plurality of mass spectra 22, e.g., using prediction
techniques described more fully below.

A number of sequence libraries can be used,
15   including, for example, the Genpept database, the GenBank
database (described in Burks, et al., "GenBank: Current status
and future directions in *Methods in Enzymology*", 183:3
(1990)), EMBL data library (described in Kahn, et al., "EMBL
Data Library," Methods in Enzymology, 183:23 (1990)), the
20   Protein Sequence Database (described in Barker, et al.,
"Protein Sequence Database," Methods in Enzymology, 1983:31
(1990), SWISS-PROT (described in Bairoch, et al., "The SWISS-
PROT protein sequence data bank, recent developments," Nucleic
Acids Res., 21:3093-3096 (1993)), and PIR-International
25   (described in "Index of the Protein Sequence Database of the
International Association of Protein Sequence Databanks (PIR-
International)" Protein Seq Data Anal. 5:67-192 (1993).

The predicted mass spectra 22 are compared 24 to
the experimentally-derived fragment spectrum 16 to identify
30   one or more of the predicted mass spectra which most closely
match the experimentally-derived fragment spectrum 16.
Preferably the comparison is done automatically by calculating
a closeness-of-fit measure for each of the plurality of
predicted mass spectra 22 (compared to the experimentally-
35   derived fragment spectrum 16). It is believed that, in
general, there is high probability that the peptide analyzed
by the tandem mass spectrometer has an amino acid sequence

identical to one of the sub-sequences taken from the protein
sequence library 20 which resulted in a predicted mass
spectrum 22 exhibiting a high closeness-of-fit with respect to
the experimentally-derived fragment spectrum 16. Furthermore,

5   when the peptide analyzed by the tandem mass spectrometer 14
was derived from a protein, it is believed there is a high
probability that the parent protein is identical or similar to
the protein whose sequence in the protein sequence library 20
includes a sub-sequence that resulted in a predicted mass

10   spectra 22 which had a high closeness-of-fit with respect to
the fragment spectrum 16. Preferably, the entire procedure
can be performed automatically using, e.g, a computer to
calculate predicted mass spectra 22 and/or to perform
comparison 24 of the predicted mass spectra 22 with the

15   experimentally-derived fragment spectrum 16.
            Fig. 3 is a flow diagram showing one method for
predicting mass spectra 22 and performing the comparison 24.
According to the method of Fig. 3, the experimentally-derived
fragment spectrum 16 is first normalized 32. According to one

20   normalization method, the experimentally-derived fragment
spectrum 16 is converted to a list of masses and intensities.
The values for the precursor ion are removed from the file.
The square root of all the intensity values is calculated and
normalized to a maximum intensity of 100. The 200 most

25   intense ions are divided into ten mass regions and the maximum
intensity is normalized to 100 within each region. Each ion
which is within 3.0 daltons of its neighbor on either side is
given the greater intensity value, if a neighboring intensity
is greater than its own intensity. Of course, other

30   normalizing methods can be used and it is possible to perform
analysis without performing normalization, although
normalization is, in general, preferred. For example, it is
possible to use maximum intensities with a value greater than
or less than 100. It is possible to select more or fewer than

35   the 200 most intense ions. It is possible to divide into more
or fewer than 10 mass regions. It is possible to make the

8

window for assuming the neighboring intensity value to be greater than or less than 3.0 daltons.

In order to generate predicted mass spectra from a protein sequence library, according to the process of Fig. 3, sub-sequences within each protein sequence are identified which have a mass which is within a tolerance amount of the mass of the unknown peptide. As noted above, the mass of the unknown peptide is known from the tandem mass spectrometer 34. Identification of candidate sub-sequences 34 is shown in greater detail in Fig. 4. In general, the process of identifying candidate sub-sequences involves summing the masses of linear amino acid sequences until the sum is either within a tolerance of the mass of the unknown peptide (the "target" mass) or has exceeded the target mass (plus tolerance). If the mass of the sequence is within tolerance of the target mass, the sequence is marked as a candidate. If the mass of the linear sequence exceeds the mass of the unknown peptide, then the algorithm is repeated, beginning with the next amino acid position in the sequence.

According to the method of Fig. 4, a variable m, indicating the starting amino acid along the sequence is initialized to 0 and incremented by 1 (36, 38). The sum, representing the cumulative mass and a variable n representing the number of amino acids thus far calculated in the sum, are initially set to 0 (40) and variable n is incremented 42. The molecular weight of a peptide corresponding to a sub-sequence of a protein sequence is calculated in iterative fashion by steps 44 and 46. In each iteration, the sum is incremented by the molecular weight of the amino acid of the next (nth) amino acid in the sequence 44. Values of the sum 44 may be stored for use in calculating fragment masses for use in predicting a fragment mass spectrum as described below. If the resulting sum is less than the target mass decremented by a tolerance 46, the value of n is incremented 42 and the process is repeated 44. A number of tolerance values can be used. In one embodiment, a tolerance value of ±0.05% of the mass of the unknown peptide was used. If the new sum is no longer less

9

than a tolerance amount below the target mass, it is then
determined if the new sum is greater than the target mass plus
the tolerance amount. If the new sum is more than the
tolerance amount in excess of the target mass, this particular
5   sequence is not considered a candidate sequence and the
process begins again, starting from a new starting point in
the sequence (by incrementing the starting point value m
(38)). If, however, the sum is not greater than the target
mass plus the tolerance amount, it is known that the sum is
10   within one tolerance amount of a target mass and, thus, that
the sub-sequence beginning with mth amino and extending to the
(m + n)th amino acid of the sequence is a candidate sequence.
The candidate sequence is marked, e.g., by storing the values
of m and n to define this sub-sequence.
15   Returning to Fig. 3, once a plurality of candidate
sub-sequences have been identified, a fragment mass spectrum
is predicted for each of the candidate sequences 52. The
fragment mass spectrum is predicted by calculating the
fragment ion masses for the type b- and y- ions for the amino
20   acid sequence. When a peptide is fragmented and the charge is
retained on the N-terminal cleavage fragment, the resulting
ion is labelled as a b-type ion. If the charge is retained on
the c-type terminal fragment, it is labelled a y-type ion.
Masses for type b- ions were calculated by summing the amino
25   acid masses and adding the mass of a proton. Type y- ions
were calculated by summing, from the c-terminus, the masses of
the amino acids and adding the mass of water and a proton to
the initial amino acid. In this way, it is possible to
calculate an m/z for each fragment. However, in order to
30   provide a predicted mass spectrum, it is also necessary to
assign an intensity value for each fragment. It might be
possible to predict, on a theoretical basis, intensity value
for each fragment. However, this procedure is difficult. It
has been found useful to assign intensities in the following
35   fashion. The value of 50.0 is assigned to each b and y ion.
To masses of 1 dalton on either side of the fragment ion, an
intensity of 25.0 is assigned. Peak intensities of 10.0 and -

17.0 and -18.0 daltons below the m/z of each b- and y- ion
location (for both NH_3 and H_2O loss), and peak intensities of
10.0 and -28.0 amu of each type b ion location (for type a-
ions).

5   Returning to Fig. 3, after calculation of predicted
m/z values and assignment of intensities, it is preferred to
calculate a measure of closeness-of-fit between the predicted
mass spectra 22 and the experimentally-derived fragment
spectrum 16. A number of methods for calculating closeness-
10   of-fit are available. In the embodiment depicted in Fig. 3, a
two-step method is used 54. The two-step method includes
calculating a preliminary closeness-of-fit score, referred to
here as S_p 56 and, for the highest-scoring amino acid
sequences, calculating a correlation function 58. According
15   to one embodiment, S_p is calculated using the following
formula:

$$S_p = \left(\sum i_m\right) * n_i * (1+\beta) * (1-\rho) / n_t \qquad (1)$$

where i_m = matched intensities, n_i = number of matched
20   fragment ions, β = type b- and y-ion continuity, ρ = presence
of immonium ions and their respective amino acids in the
predicted sequence, n_t = total number of fragment ions. The
factor, β, evaluates the continuity of a fragment ion series.
If there was a fragment ion match for the ion immediately
25   preceding the current type b- or y-ion, β is incremented by
0.075 (from an initial value of 0.0). This increases the
preliminary score for those peptides matching a successive
series of type b- and y-ions since extended series of ions of
the same type are often observed in MS/MS spectra. The factor
30   ρ evaluates the presence of immonium ions in the low mass end
of the mass spectrum. Immonium ions are diagnostic for the
presence of some types of amino acids in the sequence. If
immonium ions are present at 110.0, 120.0, or 136.0 Da ($\pm$ 1.0
Da) in the processed data file of the unknown peptide with
35   normalized intensities greater than 40.0, indicating the

presence of histidine, phenylalanine, and tyrosine respectively, then the sequence under evaluation is checked for the presence of the amino acid indicated by the immonium ion. The preliminary score, S_p, for the peptide is either augmented or depreciated by a factor of $(1 - p)$ where p is the sum of the penalties for each of the three amino acids whose presence is indicated in the low mass region. Each individual p can take on the value of -0.15 if there is a corresponding low mass peak and the amino acid is not present in the sequence, +0.15 if there is a corresponding low mass peak and the amino acid is present in the sequence, or 0.0 if the low mass peak is not present. The total penalty can range from -0.45 (all three low mass peaks present in the spectrum yet none of the three amino acids are in the sequence) to +0.45 (all three low mass peaks are present in the spectrum and all three amino acids are in the sequence).

Following the calculation of the preliminary closeness-of-fit score S_p, those candidate predicted mass spectra having the highest S_p scores are selected for further analysis using the correlation function 58. The number of candidate predicted mass spectra which are selected for further analysis will depend largely on the computational resources and time available. In one embodiment, 300 candidate peptide sequences with the highest preliminary score were selected.

For purposes of calculating the correlation function, 58, the experimentally-derived fragment spectrum is preprocessed in a fashion somewhat different from preprocessing 32 employed before calculating S_p. For purposes of the correlation function, the precursor ion was removed from the spectrum and the spectrum divided into 10 sections. Ions in each section were then normalized to 50.0. The sectionwise normalized spectra 60 were then used for calculating the correlation function. According to one embodiment, the discrete correlation between the two functions is calculated as:

12

$$R_\tau = \sum_{i=0}^{n-1} x_i y_i {+\tau} \tag{2}$$

where τ is a lag value. The discrete correlation theorem
states that the discrete correlation of two real functions x
and y is one member of the discrete Fourier transform pair

$$R_\tau \leftrightarrow X_\tau Y^* \tau \tag{3}$$

5   where X(t) and Y(t) are the discrete Fourier transforms of
x(i) and y(i) and the Y* denotes complex conjugation.
Therefore, the cross-correlations can be computed by Fourier
transformation of the two data sets using the fast Fourier
transform (FFT) algorithm, multiplication of one transform by
10   the complex conjugate of the other, and inverse transformation
of the resulting product. In one embodiment, all of the
predicted spectra as well as the pre-processed unknown
spectrum were zero-padded to 4096 data points since the MS/MS
spectra are not periodic (as intended by the correlation
15   theorem) and the FFT algorithm requires N to be an integer
power of two, so the resulting end effects need to be
considered. The final score attributed to each candidate
peptide sequence is R(0) minus the mean of the
cross-correlation function over the range $-75<t<75$. This
20   modified "correlation parameter" described in Powell and
Heiftje, Anal. Chim. Acta, Vol. 100, pp 313-327 (1978) shows
better discrimination over just the spectral correlation
coefficient R(0). The raw scores are normalized to 1.0. In
one embodiment, output 62 includes the normalized raw score,
25   the candidate peptide mass, the unnormalized correlation
coefficient, the preliminary score, the fragment ion
continuity β, the immonium ion factor ρ, the number of type b-
and y-ions matched out of the total number of fragment ions,
their matched intensities, the protein accession number, and
30   the candidate peptide sequence.
             If desired, the correlation function 58 can be used
to automatically select one of the predicted mass spectra 22

13

as corresponding to the experimentally-derived fragment spectrum 16. Preferably, however, a number of sequences from the library 20 are output and final selection of a single sequence is done by a skilled operator.

In addition to predicting mass spectra from protein sequence libraries, the present invention also includes predicting mass spectra based on nucleotide databases. The procedure involves the same algorithmic approach of cycling through the nucleotide sequence. The 3-base codons will be converted to a protein sequence and the mass of the amino acids summed in a fashion similar to the summing depicted in Fig. 4. To cycle through the nucleotide sequence, a 1-base increment will be used for each cycle. This will allow the determination of an amino acid sequence for each of the three reading frames in one pass. The scoring and reporting procedures for the search can be the same as that described above for the protein sequence database.

Depending on the computing and time resources available, it may be advantageous to employ data-reduction techniques. Preferably these techniques will emphasize the most informative ions in the spectrum while not unduly affecting search speed. One technique involves considering only some of the fragment ions in the MS/MS spectrum. A spectrum for a peptide may contain as many as 3,000 fragment ions. According to one data reduction strategy, the ions are ranked by intensity and some fraction of the most intense ions (e.g., the top 200 most intense ions) will be used for comparison. Another approach involves subdividing the spectrum into, e.g., 4 or 5 regions and using the 50 most intense ions in each region as part of the data set. Yet another approach involves selecting ions based on the probability of those ions being sequence ions. For example, ions could be selected which exist in mass windows of 57 through 186 daltons (range of mass increments for the 20 common amino acids from GLY to TRP) that contain diagnostic features of type b- or y- ions, such as losses of 17 or 18 daltons (NH_3 or H_2O) or a loss of 28 daltons (CO).

14

The techniques described above are, in general, applicable to spectra of peptides with charged states of +1 or +2, typically having a relatively short amino acid sequence. Using a longer amino acid sequence increases the probability of a unique match to a protein sequence. However, longer peptide sequences have a greater likelihood of containing more basic amino acids, and thus producing ions of higher charge state under electro-spray ionization conditions. According to one embodiment of the invention, algorithms are provided for searching a database with MS/MS spectra of highly charged peptides (+3, +4, +5, etc.). According to one approach, the search program will include an input for the charge state (N) of the precursor ion used in the MS/MS analysis. Predicted fragment ions will be generated for each charge state less than N. Thus, for peptide of +4, the charge states of +1, +2 and +3 will be generated for each fragment ion and compared to the MS/MS spectrum.

The second strategy for use with multiply charged spectra is the use of mathematical deconvolution to convert the multiply charged fragment ions to their singly charged masses. The deconvoluted spectrum will then contain the fragment ions for the multiply charged fragment ions and their singly charged counterparts.

To speed up searches of the database, a directed-search approach can be used. In cases where experiments are performed on specific organisms or specific types of proteins, it is not necessary to search the entire database on the first pass. Instead, a search sequence protein specific to a species or a class of proteins can be performed first. If the search does not provide reasonable answers, then the entire database is searched.

A number of different scoring algorithms can be used for determining preliminary closeness of fit or correlation. In addition to scoring based on the number of matched ions multiplied by the sum of the intensity, scoring can be based on the percentage of continuous sequence coverage represented by the sequence ions in the spectrum. For

15

example, a 10 residue peptide will potentially contain 9 each
of b- and y-type sequence ions. If a set of ions extends from
B_1 to B_9, then a score of 100 is awarded, but if a
discontinuity is observed in the middle of the sequence, such
5   as missing the B_5 ion, a penalty is assessed. The maximum
score is awarded for an amino acid sequence that contains a
continuous ion series in both the b and y directions.

In the event the described scoring procedures do
not delineate an answer, an additional technique for spectral
10   comparison can be used in which the database is initially
searched with a molecular weight value and a reduced set of
fragment ions. Initial filtering of the database occurs by
matching sequence ions and generating a score with one of the
methods described above. The resulting set of answers will
15   then undergo a more rigorous inspection process using a
modified full MS/MS spectrum. For the second stage analysis,
one of several spectral matching approaches developed for
spectral library searching is used. This will require
generating a "library spectrum" for the peptide sequence based
20   on the sequence ions predicted for that amino acid sequence.
Intensity values for sequence ions of the "library spectrum"
will be obtained from the experimental spectrum. If a
fragment ion is predicted at m/z 256, then the intensity value
for the ion in the experimental spectrum at m/z=256 will be
25   used as the intensity of the ion in the predicted spectrum.
Thus, if the predicted spectrum is identical to the "unknown"
spectrum, it will represent an ideal spectrum. The spectra
will then be compared using a correlation function. In
general, it is believed that the majority of computational
30   time for the above procedure is spent in the iterative search
process. By multiplexing the analysis of multiple MS/MS
spectra in one pass through the database, an overall
improvement in efficiency will be realized. In addition, the
mass tolerance used in the initial pre-filtering can affect
35   search times by increasing or decreasing the number of
sequences to analyze in subsequent steps. Another approach to
speed up searches involves a binary encryption scheme where

the mass spectrum is encoded as peak/no peak at every mass depending on whether the peak is above a certain threshold value. If intensive use of a protein sequence library is contemplated, it may be possible to calculate and store predicted mass values of all sub-sequences within a predetermined range of masses so that at least some of the analysis can be performed by table look-up rather than calculation.

Figs. 6A-6E are flow charts showing an analysis procedure according to one embodiment of the present invention. After data is acquired from the tandem mass spectrometer, as described above 602, the data is saved to a file and converted to an ASCII format 604. At this point, a preprocessing procedure is started 606. The user enters information regarding the peptide mass in the precursor ion charge state 608. Mass/intensity values are loaded from the ASCII file, with the values being rounded to unit masses 610. The previously-identified precursor ion contribution of this data is removed 612. The remaining data is normalized to a maximum intensity of 100 614. At this point, different paths can be taken. In one case, the presence of any immonium ions (H, F and Y) is noted 616 and the peptide mass and immonium ion information is stored in a datafile 618. In another route, the 200 most intense peaks are selected 620. If two peaks are within a predetermined distance (e.g., 2 amu) of each other, the lower intensity peak is set equal to a greater intensity 622. After this procedure, the data is stored in a datafile for preliminary scoring 624. In another route, the data is divided into a number of windows, for example ten windows 626. Normalization is performed within each window, for example, normalizing to a maximum intensity of 50 628. This data is then stored in a datafile for final correlation scoring 630. This ends the preprocessing phase, according to this embodiment 632.

The database search is started 634 and the search parameters and the data obtained from the preprocessing procedure (Fig. 6A) are loaded 636. A first batch of database

sequences is loaded 638 and a search procedure is run on a particular protein 640. The search procedure is detailed in Fig. 6C. As long as the end of the batch has not been reached the index is incremented 642 and the search routine is repeated 640. Once it is determined that the end of a batch has been reached 644, as long as the end of the database has not been reached, the second index 646 is incremented and a new batch of database sequences is loaded 638. Once the end of the database has been reached 628, a correlation analysis is performed 630 (as detailed in Fig. 6E), the results are printed 632 and the procedure ends 634.

When the search procedure is started 638 (Fig. 6C), an index I1 is set to zero 646 to indicate the start position of the candidate peptide within the amino acid being searched 640. A second index I2, indicating the end position of the candidate peptide within the amino acid being searched, is initially set equal to I1 and the variable Pmass, indicating the accumulated mass of the candidate peptide is initialized to zero 648. During each iteration through a given candidate peptide 650 the mass of the amino acid at position I2 is added to Pmass 652. It is next determined whether the mass thus-far accumulated (Pmass) equals the input mass (i.e., the mass of the peptide) 654. In some embodiments, this test may be performed as plus or minus a tolerance rather than requiring strict equality, as noted above. If there is equality (optionally within a tolerance) an analysis routine is started 656 (detailed in Fig. 6D). Otherwise, it is determined whether Pmass is less than the input mass (optionally within a tolerance). If not, the index I2 is incremented 658 and the mass of the amino acid at the next position (the incremented I2 position) is added to Pmass 652. If Pmass is greater than the input mass (optionally by more than a tolerance 660) it is determined whether index I1 is at the end of a protein 662. If so, the search routine exits 664. Otherwise, index I1 is incremented 666 so that the routine can start with a new start position of a candidate peptide and the search procedure returns to block 648.

18

When the analysis procedure is started 670 (Fig. 6D), data indicative of b- and y- ions for the candidate peptide are generated 672, as described above. It is determined whether the peak is within the top 200 ions 674. The peak intensity is summed and the fragmented match index is incremented 676. If previous b- or y- ions are matched 678, the β index is incremented 680. Otherwise, it is determined whether all fragment ions have been analyzed. If not, the fragment index is incremented 684 and the procedure returns to block 674. Otherwise, a preliminary score such as S_p, described above is calculated 686. If the newly-calculated S_p is greater than the lowest score 688 the peptide sequence is stored 690 unless the sequence has already been stored, in which case the procedure exits 692.

At the beginning of the correlation analysis (Fig. 6E), a stored candidate peptide is selected 693. A theoretical spectrum for the candidate peptide is created 694, correlated with experimental data 695 and a final correlation score is obtained 696, as described above. The index is incremented 697 and the process repeated from block 693 unless all candidate peptides have been scored 698, in which case the correlation analysis procedure exits 699.

The following examples are offered by way of illustration, not limitation.

Experimental

Example #1

MHC complexes were isolated from HS-EBV cells transformed with HLA-DRB*0401 using antibody affinity chromatography. Bound peptides were released and isolated by filtration through a Centricon 10 spin column. The heavy chain of glycosaparginase from human leukocytes was isolated. Proteolytic digestions were performed by dissolving the protein in 50 mM ammonium bicarbonate containing 10 mM Ca^{++}, pH 8.6. Trypsin was added in a ratio of 100/1 protein/enzyme.

Analysis of the resulting peptide mixtures was performed by LC-MS and LC-MS/MS. Briefly, molecular weights

19

of peptides were recorded by scanning Q3 or Q1 at a rate of 400 Da/sec over a mass range of 300 to 1600 throughout the HPLC gradient. Sequence analysis of peptides was performed during a second HPLC analysis by selecting the precursor ion with a 6 amu (FWHH) wide window in Q_1 and passing the ions into a collision cell filled with argon to a pressure of 3-5 mtorr. Collision energies were on the order of 20 to 50 eV. The fragment ions produced in Q_2 were transmitted to Q_3 and a mass range of 50 Da to the molecular weight of the precursor ion was scanned at 500 Da/sec to record the fragment ions. The low energy spectra of 36 peptides were recorded and stored on disk. The genpept database contains protein sequences translated from nucleotide sequences. A text search of the database was performed to determine if the sequences for the peptide amino acid sequences used in the analysis were present in the database. Subsequently, a second database was created from the whole database by appending amino acid sequences for peptides not included.

The spectrum data was converted to a list of masses and intensities and the values for the precursor ion were removed from the file. The square root of all the intensity values was calculated and normalized to a maximum intensity of 100.0. All ions except the 200 most intense ions were removed from the file. The remaining ions were divided into 10 mass regions and the maximum intensity normalized to 100.0 within each region. Each ion within 3.0 daltons of its neighbor on either side was given the greater intensity value, if the neighboring intensity was greater than its own intensity. This processed data was stored for comparison to the candidate sequences chosen from the database search. The MS/MS spectrum was modified in a different manner for calculation of a correlation function. The precursor ion was removed from the spectrum and the spectrum divided into 10 equal sections. Ions in each section were then normalized to 50.0. This spectrum was used to calculate the correlation coefficient against a predicted MS/MS spectrum for each amino acid sequence retrieved from the database.

Amino acid sequences from each protein were
generated by summing the masses, using average masses for the
amino acids, of the linear amino acid sequence from the amino
terminus (n). If the mass of the linear sequence exceeded the
mass of the unknown peptide, then the algorithm returned to
the amino terminal amino acid and began summing amino acid
masses from the n+1 position. This process was repeated until
every linear amino acid sequence combination had been
evaluated. When the mass of the amino acid sequence was
within ±0.05% (minimum of ±1 Da) of the mass of the unknown
peptide, the predicted m/z values for the type b- and y-ions
were generated and compared to the fragment ions of the
unknown sequence. A preliminary score (S_p) was generated and
the top 300 candidate peptide sequences with the highest
preliminary score were ranked and stored. A final analysis of
the top 300 candidate amino acid sequences was performed with
a correlation function. Using this function a theoretical
MS/MS spectrum for the candidate sequence was compared to the
modified experimental MS/MS spectrum. Correlation
coefficients were calculated, ranked and reported. The final
results were ranked on the basis of the normalized correlation
coefficient.

The spectrum shown in Fig. 5 was obtained by
LC-MS/MS analysis of a peptide bound to a DRB*0401 MHC class
II complex. A search of the genpept database containing
74,938 protein sequences identified 384,398 peptides within a
mass tolerance of ±0.05% (minimum of ±1Da) of the molecular
weight of this peptide. By comparing fragment ion patterns
predicted for each of these amino acid sequences to the
pre-processed MS/MS spectra and calculating a preliminary
score, the number of candidate sequences was cutoff at 300. A
correlation analysis was then performed with the predicted
MS/MS spectra for each of these sequences and the modified
experimental MS/MS spectrum. The results of the search
through the genpept database with the spectrum in Fig. 5 are
displayed in Table 1. Two peptides of similar sequence,
DLRSWTAADAAQISK [Seq. ID No. 1], DLRSWTAADAAQISQ [Seq. ID No.

21

2], were identified as the highest scoring sequences (C_n values). Their correlation coefficients are identical so their rankings in Table 1 are arbitrary. The amino acid sequence DLRSWTAADAAQISK [Seq. ID No. 1] occurs in five proteins in the genpept database while the sequence DLRSWTAADAAQISQ [Seq. ID No. 2] occurs in only one. The top three sequences appear in immunologically related proteins and the rest of the proteins appear to have no correlation to one another. A second search using the same MS/MS spectrum was performed with the *Homo sapiens* subset of the genpept database to compare the results. These data are presented in Table 2. In both searches the correct sequence tied for the top position. Both amino acid sequences have identical correlation coefficients, C_n, although the sequences differ by Lys and Gln at the C-terminus. These two amino acids have the same nominal mass and would be expected to produce similar MS/MS spectra. The sum of the normalized fragment ion intensities, I_m, for the matched fragment ions for the two peptides are different with the correct sequence having the greater value. The correct sequence also matched an additional fragment ion in the preliminary scoring procedure identifying 70% of the predicted fragment ions for this amino acid sequence in the pre-processed spectrum. These matches are determined as part of the preliminary scoring procedure.

25

Table 1

No	Mass	C_n	C	S_p	β	ρ	Ions	I_m	Accession No	No of Proteins	Sequence
1	1734.90	1.000	4.203	1809.9	1.125	0.000	20/30	1277.61	X17273	+5	DLRSWTAADTAAQISK
2	1734.86	1.000	4.203	2020.1	1.200	0.000	21/30	1311.79	M84380	+2	DLRSWTAADTAAQISQ
3	1733.91	0.893	3.752	1440.8	0.975	0.000	19/30	1151.87	X60251		DLRSWTAADTAAQVTQ
4	1733.96	0.727	3.057	606.1	0.525	0.000	14/30	851.71	m90516		VIFLEDDDVAAVVDGR
5	1733.96	0.685	2.881	1012.5	0.750	0.000	16/28	1012.47	D13262	+1	TDMEELLAGIWQDVL
6	1734.94	0.684	2.874	805.8	0.675	0.000	16/32	962.12	S58174		RCLTASTLGLTGNVNVN
7	1743.96	0.683	2.869	763.4	0.600	0.000	15/30	954.28	X17149		GSSLMAEDNLELAVRK
8	1733.97	0.767	2.839	790.1	0.675	0.000	17/36	998.90	M19085		SGILTLSGGLSSNAEVTAG
9	1733.93	0.675	2.836	682.9	0.525	0.000	13/26	895.55	M29146	+1	YPSKQINELWENVL
10	1734.90	0.668	2.809	590.8	0.525	0.000	15/36	929.73	M84615	+2	AKSGEANVTSATGTIGGTI
11	1733.99	0.663	2.786	889.5	0.750	0.000	16/32	1016.53	D13297		GKENNKSAASSKGKITL
12	1734.94	0.648	2.724	459.8	0.450	0.000	13/30	731.82	M32066		MLLDAIKGGSSDLHFE
13	1735.89	0.639	2.685	1461.3	0.900	0.000	19/30	1214.40	M28242	+1	FTPESVSRLLEKISAG
14	1733.95	0.636	2.672	614.5	0.600	0.000	13/28	827.23	J05110		SKPKYNNEVEAKLDV
15	1733.99	0.632	2.658	501.7	0.525	0.000	12/30	865.03	M24378		MTNTNMHAGVNNSQSQ
16	1733.93	0.628	2.641	563.5	0.300	0.000	14/30	928.90	M76547		KTDSNGNIKLDCPSLK
17	1734.91	0.626	2.629	580.6	0.600	0.000	13/28	781.63	M30149	+5	ENLLNERGYGKLVD
18	1734.99	0.622	2.613	756.6	0.525	0.000	17/34	992.25	M33962		SGSLSPEHGPIVVHCSAG
19	1734.99	0.619	2.600	625.6	0.450	0.000	14/30	924.55	K02670		LFVLLAVFIAGLMIGR
20	1734.91	0.617	2.593	563.2	0.600	0.000	12/26	762.63	M63585	+1	DILDKRLFWAQDGR
21	1735.86	0.617	2.592	491.3	0.525	0.000	13/30	743.39	X67813		AQLISAYSLVDPEKAK
22	1735.05	0.615	2.585	472.5	0.600	0.000	12/28	689.08	J04120	+2	QELVIHIGWIISNNP
23	1734.94	0.615	2.583	548.0	0.450	0.000	13/30	872.19	X54240		VDESINEDNSVVSLSQ
24	1734.63	0.609	2.561	493.4	0.525	0.000	12/28	754.92	X17637		DVLANAYRISRQEAQ
25	1734.99	0.603	2.534	602.6	0.450	0.000	14/28	831.23	D90034		CEDSLDERKIKGVIE

Table 1 (continued)

Mass = calculated candidate peptide mass, C_n = normalized correlation coefficient, C = correlation coefficient, S_p = preliminary score, β = fragment ion continuity, Ions = number of type b- and y-ions matched out of the total number of fragment ions for candidate sequence, Im = matched fragment intensities, Accession No = genpept accession number, and the candidate peptide sequence, mass = 1734.90, frag tolerance = 0.75, mass tolerance = 1.000.

Table 2

No	Mass	c_n	C	S_p	β	ρ	Ions	I_m	Accession No	No of Proteins	Sequence
1	1734.90	1.000	4.203	1809.9	1.125	0.000	20/30	1277.61	X17273	+5	DLRSWTAADTAAQISK
2	1734.86	1.000	4.203	2020.1	1.200	0.000	21/30	1311.79	M84380		DLRSWTAADTAAQISQ
3	1733.96	0.727	3.057	606.1	0.525	0.000	14/30	851.71	M90516		VIFLEDDVAAVVDGR
4	1733.96	0.598	2.514	642.1	0.600	0.000	15/32	856.16	X54637		GSSGRNPQASLFGKKAK
5	1733.99	0.577	2.424	465.9	0.275	0.000	12/28	790.70	X02598		CKLQSGTHCLWTDQL
6	1733.93	0.558	2.347	674.2	0.450	0.000	15/32	991.85	Z18330	+5	MTQSPATLSVSPGERAT
7	1733.93	0.554	2.329	300.2	0.225	0.000	10/28	686.18	Z22658		PYVGKELNMIIMLpD
8	1734.12	0.553	2.323	447.6	0.375	0.000	13/32	801.32	X02763		SSGILSRSSVGPCIRSQ
9	1734.96	0.549	2.307	375.1	0.225	0.000	12/30	765.57	J03259		FPVDTTELVLTGNNLT
10	1734.94	0.545	2.292	544.9	0.375	0.000	14/30	849.24	M18728		PASFAWQDDAVISISQ
11	1735.89	0.543	2.281	590.0	0.450	0.000	13/28	876.39	J03202	+1	ETLENEANNIKMEAE
12	1735.86	0.542	2.276	646.1	0.450	0.000	15/30	891.21	X51592	+2	KCEIQAALEEAEASLE
13	1734.91	0.537	2.259	574.0	0.450	0.000	13/26	791.73	X63575	+2	MTNSDFYSKNQRNE
14	1734.83	0.534	2.243	500.9	0.525	0.000	13/30	758.05	X15875	+1	LTQMADQSTEPALSQI
15	1733.93	0.532	2.235	514.3	0.600	0.000	13/30	741.72	S79307		LRLSSVTAADTAVYYC
16	1733.97	0.526	2.210	377.1	0.300	0.000	12/28	676.80	X12530	+2	SEKNSPSTQYCYSIQ
17	1735.86	0.525	2.206	518.3	0.525	0.000	15/34	770.42	M25668		TAAGGESALAPSVFKQAK
18	1733.96	0.525	2.205	360.3	0.225	0.000	12/30	735.30	M81830		SSGIRVGSSKRKKSEK
19	1734.99	0.524	2.202	376.7	0.300	0.000	12/30	724.46	K03456	+2	SEGATPQDLNMLNIV
20	1733.99	0.522	2.195	863.6	0.675	0.000	15/28	962.41	M92269	+1	ECLKRQKDRGGDISQ
21	1733.94	0.520	2.187	603.4	0.525	0.000	15/36	949.64	M85289		AKGSVYIGGAPDVATLTG
22	1734.94	0.520	2.187	540.6	0.450	0.000	13/28	802.95	M17081	+4	KTDIKVVDRDSEEAE
23	1734.85	0.516	2.167	526.1	0.750	0.000	13/28	647.52	L07217	+5	IIRSENFTNNAKTII
24	1734.99	0.515	2.166	437.6	0.450	0.000	12/30	754.47	X70848	+3	VDEKIFCCHGGLSPDL
25	1735.67	0.510	2.143	532.9	0.450	0.000	13/30	848.14	M95929		SGSDTPQQDNDQLNSE

25

Example #2

To examine the complexity of the mixture of
peptides obtained by proteolysis of the total proteins from *S. cerevisiae* cells, 10^8 cells were grown and harvested. After
lysis, the total proteins were contained in ~9 mL of solution.
A 0.5 mL aliquot was removed for proteolysis with the enzyme
trypsin. From this solution two microliters were directly
injected onto a micro-LC (liquid chromatography) column for MS
analysis. In a complex mixture of peptides it is conceivable
that multiple peptide ions may exist at the same m/z and
contribute to increased background, complicating MS/MS
analysis and interpretation. To test the ability to obtain
sequence information by MS/MS from these complex mixtures of
peptides, ions from the mixture were selected with on-line
MS/MS analysis. In no case were the spectra contaminated with
fragment ions from other peptides. A partial list of the
identified sequences is presented in Table 3.

Table 3

S. cerevisiae Protein	Seq. ID No.	Amino acid Sequence
enolase	3	DPFAEDDWEAWSH
hypusine containing protein HP2	4	APEGELGDSLQTAFDEGK
phosphoglycerate kinase	5	TGGGASLELLEGK
BMH1 gene product	6	QAFDDAIAELDTLSEESYK
pyruvate kinase	7	IPAGWQGLDNGPSER
phosphoglycerate kinase	8	LPGTDVDLPALSEK
hexokinase	9	IEDDPFENLEDTDDDFQK
enolase	10	EEALDLIVDAIK
enolase	11	NPTVEVELTTEK

The MS/MS spectra presented in Table 1 were
interpreted using the described database searching method.
This method serves as a data pre-filter to match MS/MS spectra
to previously determined amino acid sequences. Pre-filtering
the data allows interpretation efforts to be focused on
previously unknown amino acid sequences. Results for some of
the MS/MS spectra are shown in Table 4. No pre-assigning of
sequence ions or manual interpretation is required prior to
the search. However, the sequences must exist in the

26

database. The algorithm first pre-processed the MS/MS data
and then compared all the amino acid sequences in the database
within ±1 amu of the mass of the precursor ion of the MS/MS
spectrum. The predicted fragmentation patterns of the amino
acid sequences within the mass tolerance were compared to the
experimental spectrum. Once an amino acid sequence was within
this mass tolerance, a final closeness-of-fit measure was
obtained by reconstructing the MS/MS spectra and performing a
correlation analysis to the experimental spectrum. Table 4
lists a number of spectra used to test the efficacy of the
algorithm.

        The computer program described above has been
modified to analyze the MS/MS spectra of phosphorylated
peptides. In this algorithm all types of phosphorylation are
considered such as Thr, Ser, and Tyr. Amino acid sequences
are scanned in the database to find linear stretches of
sequence that are multiples of 80 amu below the mass of the
peptide under analysis. In the analysis each putative site of
phosphorylation is considered and attempts to fit a
reconstructed MS/MS spectrum to the experimental spectrum are
made.

Table 4

List of results obtained searching genpept and species specific databases using MS/MS spectra for the respective peptides.

No.	Mass	Amino Acid Sequence of Peptides used in the Search	Seq. ID No.	Genpept Database	Genpept Database[3]	Species Specific
1	1734.9	DLRSWTAADTAAQISQ	12	1	1	1
2	1749	DLRSWTAADTAAQITQ	13	1	1	1
3	1186.5	MATPLLMQALP	14	-	-	13
4	1317.7	MATPLLMQALP	14	61*	61	17
5	1571.6	EGVNDNEEGFFSAR[1,2]	15	1*	1	1
6	1571.6	EGVNDNEEGFFSAR[1,2]	15	1*	1	1
7	1297.5	DRVYIHPFHL(+2)	16	1	1	1
8	1297.5	DRVYIHPFHL(+2)	16	2	2	2
9	1297.5	DRVYIHPFHL(+3)	16	1	1	1
10	1593.8	VEADVAGHGQDILIR[2]	17	1	1	1
11	1393.7	HGVTVLTALGAILK[2]	18	1	1	1
12	1741.8	HSGQAEGYSYTDANIK[2]	19	1	1	1
13	848.8	HSGQAEGY[2](+1)	20	1	1	1
14	723.9	MAFGGLK[2,3](+1)	21	-	-	-
15	636.8	GATLFK[2](+1) [QATLFG, KTLFK]	22	1*	1*	6
16	524.6	TEFK(+1)	23	1*	1*	5
17	1251.4	DRNDLLTYLK[1,2]	24	5*	5	1
18	1194.4	VLVLDTDYKK[2]	25	6*	6	2
19	700.7	CRGDSY[1] (CGRDSY)	26	3*	1	1
20	700.7	CRGDSY[1](+1)	26	-	-	7
21	764.9	KGATLFK[2]	27	3	3	1
22	1169.3	TGPNLHGLFGR	28	1	1	1
23	1047.2	DRVYIHPF	29	-	-	7
24	1139.3	TLLVGESATTF(+1)	30	1	1	1
25	1189.4	RNVIPDSKY	31	1	1	1
26	613.7	SSPLPL(+1)	32	2	4	2
27	1323.5	LARNCQPNYW(C=161.17)	33	1	1	1
28	2496.7	AQSMGFINEDLSTSAQALMSDW	34	1	1	1
29	1551.8	VTLIHPIAMDDGLR	35	3	3	1
30	1803.0	GGDTVTLNETDLTQIPK	36	2	2	1
31	1172.4	VGEEVEIVGIK	37	1	1	1
32	2148.5	GWQVPAFTLGGEATDIVVMR	38	1	1	1
33	2553.0	VASISLPTSCASAGTQCLISGWGNTK[1]	39	-	1	1
34	1154.3	SSGTSYPDVLK[1]	40	-	3	1
35	1174.5	TLNNDIMLIK	41	1	1	1
36	2274.6	SIVHPSYNSNTLNNDIMLIK[1]	42	-	2	1

1. not present in the genpept database
2. sequence appended to the human database, not originally in human database
3. amino acid sequences added to database
(-) not in the top 100 answers
* peptide of similar sequence identified

Example #3

Much of the information generated by the genome projects will be in the form of nucleotide sequences. Those stretches of nucleotide sequence that can be correlated to a gene will be translated to a protein sequence and stored in a

specific database (genpept). The un-translated nucleotide
sequences represent a wealth of data that may be relevant to
protein sequences. The present invention will allow searching
the nucleotide database in the same manner as the protein

5 sequence databases. The procedure will involve the same
algorithmic approach of cycling through the nucleotide
sequence. The three-base codon will be converted to a protein
sequence and the mass of the amino acids summed. To cycle
through the nucleotide sequence, a one-base increment will be

10 used for each cycle. This will allow the determination of an
amino acid sequence for each of the three reading frames in
one pass. For example, an MS/MS spectrum is generated for the
sequence Asp-Leu-Arg-Ser-Trp-Thr-Ala [Seq. ID No. 43]
((M+H)+=848) the algorithm will search the nucleotide sequence

15 in the following manner.

Seq. ID No.

Nucleotide sequence from the database.
nucleotides GCG AUC UCC GGU CUU GGA CUG CUC   44
First pass through the sequence.

20 nucleotides GCG AUC UCC GGU CUU GGA CUG CUC   Mass   44
amino acids Ala Ile Ser Gly Leu Gly Leu Leu   743   45
Second pass through the sequence.
nucleotides G   CGA UCU CCG GUC UUG GAC UGC UC   Mass   44
amino acids   Arg Ser Pro Val Leu Gly Leu   741   46

25 Third pass through the sequence.
nucleotides GC GAU CUC CGG UCU UGG ACU GCU C   Mass   44
amino acids   Asp Leu Arg Ser Trp Thr Ala   848   43
Fourth pass through the sequence.
nucleotides GCG AUC UCC GGU CUU GGA CUG CUC   Mass   44

30 amino acids   Ile Ser Gly Leu Gly Leu Leu   672   45
As the sequence of amino acids match the mass of the peptide
the predicted sequence ions will be compared to the MS/MS
spectrum. From this point on the scoring and reporting
procedures for the search will be the same as for a protein

35 sequence database.
        In light of the above description, a number of
advantages of the present invention can be seen. The present
invention permits correlating mass spectra of a protein,
peptide or oligonucleotide with a nucleotide or protein

40 sequence database in a fashion which is relatively accurate,

rapid, and which is amenable to automation (i.e., to operation without the need for the exercise of human judgment). The present invention can be used to analyze peptides which are derived from a mixture of proteins and thus is not limited to analysis of intact homogeneous proteins such as those generated by specific and known proteolytic cleavage.

A number of variations and modifications of this invention can also be used. The invention can be used in connection with a number of different proteins or peptide sources and it is believed applicable to any analysis using mass spectrometry and proteins. In addition to the examples described above, the present invention can be used for, for example, monitoring fermentation processes by collecting cells, lysing the cells to obtain the proteins, digesting the proteins, e.g. in an enzyme reactor, and analyzing by Mass spectrometry as noted above. In this example, the data could be interpreted using a search of the organism's database (e.g., a yeast database). As another example, the invention could be used to determine the species of organism from which a protein is obtained. The analysis would use a set of peptides derived from digestion of the total proteins. Thus, the cells from the organism would be lysed, the proteins collected and digested. Mass spectrometry data would be collected with the most abundant peptides. A collection of spectra (e.g., 5 to 10 spectra) would be used to search the entire database. The spectra should match known proteins of the species. Since this method would use the most abundant proteins in the cell, it is believed that there is a high likelihood the sequences for these organisms would be sequenced and in the database. In one embodiment, relatively few cells could be used for the analysis (e.g., on the order of $10^4 - 10^5$).

For example, methods of the invention can be used to identify microorganisms, cell surface proteins and the like. For identifying microorganisms, the procedure can employ tandem mass spectra obtained from peptides produced by proteolytic digestion of the cellular proteins. The complex

mixture of peptides produced is subjected to separation by
HPLC on-line to a tandem mass spectrometer. As peptides elute
off the column tandem mass spectra are obtained by selecting a
peptide ion in the first mass analyzer, sending it into a
collision cell, and recording the mass-to-charge (m/z) ratios
of the resulting fragment ions in the second mass analyzer.
This process is performed over the course of the HPLC analysis
and produces a large collection of spectra (e.g., from 10 to
200 or more). Each spectrum represents a peptide derived the
microorganism's protein (gene) pool and thus the collection
can be used to develop one or more family, genus, species,
serotype or strain-specific markers of the microorganism, as
desired.

     The identification of the microorganism is performed
using one of at least three software related techniques. In a
first technique, a database search, the tandem mass spectra
are used to search protein and nucleotide databases to
identify an amino acid sequence which is represented by the
spectrum. Identification of the organism is achieved when a
preponderance of spectra obtained in the mass spectrometry
analysis match to proteins previously identified as coming
from a particular organism. Means for searching databases in
this fashion are as described hereinabove.

     In a second technique a library search can be
performed, such as if no solid matches are observed using the
database search described above. In this approach the data
set is compared to a pre-defined library of spectra obtained
from known organisms. Thus, initially a library of peptide
spectra is created from known microorganisms. The library of
tandem mass spectra for micro-organisms can be constructed by
any of several methods which employ LC-MS/MS. The methods can
be used to vary the location cellular proteins are obtained
from, and the amount of pre-purification employed for the
resulting peptide mixture prior to LC-MS/MS analysis. For
example, intact cells can be treated with a proteolytic enzyme
such as trypsin, chymotrypsin, endoproteinase Glu-C,
endoproteinase Lys-C, pepsin, etc. to digest the proteins

exposed on the cell surface. Pre-treatment of the intact
cells with one or more glycosidases can be used to remove
steric interference that may be created by the presence of
carbohydrates on the cell surface. Thus, the pre-treatment
5   with glycosidases may be used to obtain higher peptide yields
during the proteolysis step. A second method to prepare
peptides involves rupturing the cell membranes (e.g., by
sonication, hypo-osmotic shock, freeze-thawing, glass beads,
etc.) and collecting the total proteins by precipitation,
10   e.g., using acetone or the like. The proteins are resuspended
in a digestion buffer and treated with a protease such as
trypsin, chymotrypsin, endoproteinase glu-C, endoproteinase
lys-C, etc. to create a mixture of peptides. Partial
simplification of this mixture of peptides, such as by
15   partitioning the mixture into acid and basic fractions or by
separation using strong cation exchange chromatography, leads
to several pools of peptides which can then be used in the
mass spectrometry process. The peptide mixtures are analyzed
by LC-MS/MS, creating a large set of spectra, each
20   representing a unique peptide marker of the organism or cell
type.
        The data are stored in the library in any of a
variety of means, but conveniently in three sections, wherein
one section is the peptide mass determined from the spectrum,
a second section is information related to the organism,
25   species, growth conditions, etc., and a third section contains
the mass/intensity data. The data can be stored in a variety
of formats, conveniently an ASCII format or in a binary
format.
        To perform the library search spectra are compared
30   by first determining whether the mass of the peptide is within
a preset mass tolerance (typically about ± 1-3 amu) of the
library spectrum; a cross-correlation function as described
hereinabove is used to obtain a quantitative value of the
similarity or closeness-of-fit of the two spectra. The
35   process is similar to the database searching algorithm except
a spectrum is not reconstructed for the amino acid sequence.

To provide a set of comparison spectra the tandem mass spectrum can be used to search a small (e.g., ~100 protein sequences) randomly generated sequence database. This provides a background against which similarity is compared and to generate a normalized score.

A third related technique for identifying a microorganism or cell involves de novo interpretation to determine a set of amino acid sequences that have the same mass as the peptide represented by the spectrum. The set of amino acid sequences is limited by using the spectral pre-processing equation 1, above, to rank the sequences. This set of amino acid sequences then serves as the database for use in the searching method described hereinabove. An amino acid sequence is thereby derived for a tandem mass spectrum that is not contained in the organized databases. By using phylogenetic analysis of the determined amino acid sequences they can be placed within a species, genus or family and a classification of the microorganism is thereby accomplished.

The methodology described above has applications in addition to identifying microorganisms. For example, cDNA sequencing can be carried out using conventional means to obtain partial sequences of genes expressed in particular cell lines, tissue types or microorganisms. This information then serves as the database for the subsequent analyses. The approach described above for the digesting proteins exposed on the cell surface by enzymatic digestion can be used to generate a collection of peptides for LC-MS/MS analysis. The resulting spectra are used to search the nucleotide sequences in all 6 reading frames to match amino acid sequences to the MS/MS spectra. The amino acid sequences identified represent regions of the cell surface proteins exposed to the extracellular space. This method provides at least two additional pieces of information not directly obtainable from cDNA sequencing. First, the spectra identify the proteins residing on the membrane of the cells. Secondly, sidedness information is obtained about the folding of the proteins on the cell surface. The peptide sequences matched to the

nucleotide sequence information identifies those segments of the protein sequence exposed extracellularly.

The methods can also be used to interpret the MS/MS spectra of carbohydrates. In this method the carbohydrate(s) of interest is subjected to separation by HPLC on-line to a tandem mass spectrometer as with the peptides. The carbohydrates can be obtained from a complex mixture of carbohydrates or obtained from proteins, cells, etc. by chemical or enzymatic release. Tandem mass spectra are obtained by selecting a carbohydrate ion in the first mass analyzer, sending it into a collision cell, and recording the mass-to-charge (m/z) ratios of the resulting fragment ions in the second mass analyzer. This process is performed over the course of the HPLC analysis and produces a large collection of spectra (e.g., from 10 to 200 or more). The fragmentation patterns of the carbohydrate structures contained in the database can be predicted and a theoretical representation of the spectra can be compared to the pattern in the tandem mass spectrum by using the method described hereinabove. The carbohydrate structures analyzed by tandem mass spectrometry can thereby be identified. These methods can thus be used for characterization of the carbohydrate structures found on proteins, cell surfaces, etc.

The present invention can be used in connection with diagnostic applications, such as described above and in Example 2. Another example involves identifying virally infected cells. Success of such an approach is believed to depend on the relative abundance of the viral proteins versus the cellular proteins, at least using present equipment. If an antibody were produced to a specific region of a protein common to certain pathogens, the mixture of proteins could be partially fractionated by passing the material over an immunoaffinity column. Bound proteins are eluted and digested. Mass spectrometry generates the data to search a database. One important element is finding a general handle to pull proteins from the cell. This approach could also be used to analyze specific diagnostic proteins. For example, if

a certain protein variant is known to be present in some form of cancer or genetic disease, an antibody could be produced to a region of the protein that does not change. An immunoaffinity column could be constructed with the antibody to isolate the protein away from all the other cellular proteins. The protein would be digested and analyzed by tandem mass spectrometry. The database of all the possible mutations in the protein could be maintained and the experimental data analyzed against this database.

One possible example would be cystic fibrosis. This disease is characterized by multiple mutations in CFTR protein. One mutation is responsible for about 70% of the cases and the other 30% of the cases result from a wide variety of mutations. To analyze these mutations by genetic testing would require many different analyses and probes. In the assay described above, the protein would be isolated and analyzed by tandem mass spectrometry. All the mutations in the protein could be identified in an assay based on structural information. The database used would preferably describe all the known mutations. Implementation of this approach is believed to involve significant difficulties. The amount of protein required could be so large that it would be impractical to obtain from a patient. This problem may be overcome as the sensitivity of mass spectrometry improves in the future. A protein such as CFTR is a transmembrane protein, which are typically very difficult to manipulate and digest. The approach described could be used for any diagnostic protein. The data would be highly specific and the data analysis essentially automated.

It is believed that the present invention can be used with any size peptide. The process requires that peptides be fragmented and there are methods for achieving fragmentation of very large proteins. Some such techniques are described in Smith et al., "Collisional Activation and Collision-Activated Dissociation of Large Multiply Charged Polypeptides and Proteins Produced by Electrospray Ionization" J. Amer. Soc. Mass Spect. I: 53-65 (1990). The present method

35

can be used to analyze data derived from intact proteins, in
that there is no theoretical or absolute practical limit to
the size of peptides that can be analyzed according to this
invention. Analysis using the present invention has been
5   performed on peptides at least in the size range from about
400 amu (4 residues) to about 2500 amu (26 residues).

In described embodiments candidate sub-sequences are
identified and fragment spectra are predicted as they are
needed, at the time of doing the analysis. If sufficient
10   computational resources and storage facilities are available
to perform some or all of the calculations needed for
candidate sequence identification (such as calculation of sub-
sequence masses) and/or spectra prediction (such as
calculation of fragment masses), storage of these items in a
15   database can be employed so that some or all of these items
can be looked up rather than calculated each time they are
needed.

While the present invention has been described by
way of the preferred embodiment and certain variations and
20   modifications, other variations and modifications of the
present invention can also be used, the invention being
described by the following claims.

36

WHAT IS CLAIMED IS:

1.   A method for correlating a peptide fragment
mass spectrum with amino acid sequences derived from a
database of sequences, comprising:
        storing data representing a first mass spectrum of a
plurality of fragments of at least a first peptide;
        calculating a plurality of predicted mass spectra of
at least a portion of a plurality of said sequences in said
database of sequences; and
        calculating at least a first measure for each of
said plurality of predicted mass spectra, said first measure
being an indication of the closeness-of-fit between said first
mass spectrum and each of said plurality of mass spectra.

2.   A method, as claimed in claim 1, wherein said
first mass spectrum is provided from a tandem mass
spectrometer device.

3.   A method, as claimed in claim 2, wherein the
tandem mass spectrometer is one of a triple quadrupole mass
spectrometer, a Fourier-transform cyclotron resonance mass
spectrometer, a tandem time-of-flight mass spectrometer and a
quadrupole ion trap mass spectrometer.

4.   A method, as claimed in claim 1, wherein said
database of sequences is a database of amino acid sequences of
a plurality of proteins.

5.   A method, as claimed in claim 1, wherein said
database of sequences is a nucleotide database.

6.   A method, as claimed in claim 1, further
comprising selecting a first plurality of sub-sequences from
said database of sequences, wherein said step of calculating a
plurality of predicted mass spectra includes calculating at
least one predicted mass spectrum for each of said selected
first plurality of sub-sequences.

7.   A method, as claimed in claim 1, wherein said step of calculating a first measure includes selecting those values from said first mass spectrum having an intensity greater than a predetermined threshold.

8..   A method, as claimed in claim 1, further comprising normalizing said first spectrum prior to said step of calculating at least a first measure.

9.   A method, as claimed in claim 1, wherein said step of calculating a plurality of predicted mass spectra includes calculating predicted mass spectra for only a portion of said sequence database.

10.   A method, as claimed in claim 9, wherein said first peptide is derived from a protein which is obtained from a first organism and wherein said protein of said sequence database is the portion containing sequences for proteins found in said first organism.

11.   A method, as claimed in claim 2 wherein a first mass spectrometer of said tandem mass spectrometer device is used to separate-out a component having a first mass, an activation device of said mass spectrometer device is used to fragment said first component and a second mass spectrometer of said tandem mass spectrometer device is used provide said first mass spectrum.

12.   A method, as claimed in claim 1, wherein said first peptide is isolated by chromatography.

13.   A method, as claimed in claim 1, wherein said data representing said first mass spectrum includes a plurality of mass-charge pairs.

38

1   14. A method, as claimed in claim 1, wherein said
2 step of calculating predicted mass spectra comprises:
3   deriving a plurality of masses from portions of said
4 plurality of sequences, each mass equal to the mass of a
5 peptide fragment which corresponds to a portion of a sequence
6 in said plurality of sequences;
7   selecting those masses, among said plurality of
8 masses, which are within a predetermined mass tolerance of the
9 mass of said first peptide and storing an indication of which
10 portion of which sequence each of said selected masses
11 corresponds to, to provide a plurality of candidate sequence
12 portions; and
13   calculating a plurality of mass-charge pairs for
14 each of said candidate sequence portions, each of said mass-
15 charge pairs having a mass substantially equal to the mass of
16 a peptide fragment corresponding to a portion of one of said
17 candidate sequence portions.

1   15. A method, as claimed in claim 1, wherein said
2 first measure comprises a correlation coefficient.

1   16. A method, as claimed in claim 1, wherein said
2 step of calculating a first measure comprises:
3   calculating a preliminary score for each of said
4 plurality of candidate sequence portions;
5   identifying a plurality of primary candidate
6 portions which have a preliminary score which is greater than
7 at least one candidate sequence which is not identified as a
8 primary candidate portion; and
9   calculating a correlation coefficient for each of
10 said primary candidate portions.

1   17. A method, as claimed in claim 8, wherein each
2 of said plurality of mass spectra and said first mass spectrum
3 includes a plurality of mass-charge pairs, each mass-charge
4 pair having an intensity value, and further comprising the
5 step of identifying, for each of said plurality of mass

6   spectra, a set of matched fragments which have less than a
7   predetermined difference from corresponding fragments in said
8   first mass spectrum; and
9   wherein said preliminary score is the number of
10   fragments of a predicted mass spectrum in said set of matched
11   fragments multiplied by the sum of the intensity values for
12   the mass-charge pairs corresponding to said matched fragments.

1   18.   A method for interpreting the mass spectrum of
2   an oligonucleotide comprising:
3   providing a library of nucleotide sequences;
4   storing, in a database, a plurality of nucleotide
5   sub-sequences from said library, said plurality including all
6   sequences smaller than n-mers;
7   storing data representing a first mass spectrum of a
8   plurality of fragments of said oligonucleotide;
9   calculating predicted mass spectra for each of said
10   plurality of nucleotide sub-sequences; and
11   calculating at least a first closeness-of-fit
12   measure for each of said predicted mass spectra, with respect
13   to said first mass spectrum.

1   19.   A method, as claimed in claim 18, wherein n is
2   10.

1   20.   A method for determining whether a peptide in a
2   mixture of proteins is homologous to a portion of any of a
3   plurality of proteins specified by an amino acid sequence in a
4   database of sequences, comprising:
5   using a tandem mass spectrometer to receive a
6   plurality of peptides obtained from said mixture of proteins,
7   to select at least a first peptide from said mixture of
8   peptides, to fragment said first peptide and to generate a
9   peptide fragment mass spectrum;
10   storing data representing said first mass spectrum;
11   and

40

12   correlating said mass spectrum with an amino acid
13 sequence in said database of sequences, to determine the
14 correspondence of a protein specified in said sequence
15 database with a protein in said mixture of proteins.

1   21. A method, as claimed in claim 20, wherein said
2   step of correlating includes predicting at least one mass
3   spectrum from said amino acid sequence.

1   22. A method according to claim 20 wherein the
2   mixture of proteins is obtained from a cell or microorganism
3   to be identified.

1   23. A method according to claim 22, wherein the
2   mixture of proteins is obtained by proteolytic digestion of
3   cellular proteins.

1   24. The method of claim 23, wherein the cellular
2   proteins are extracellular.

1   25. A method for identifying an organism of
2   interest by determining whether a mass spectrum or a plurality
3   of mass spectra of peptides obtained from the organism or
4   components thereof to be identified is contained in a library
5   of spectra of known organisms, comprising:
6   using a tandem mass spectrometer to receive a
7   plurality of peptides obtained from a mixture of proteins
8   obtained from said organism to be identified, to select at
9   least a first peptide from said plurality of peptides, to
10 fragment said first peptide and to generate a peptide fragment
11 mass spectrum;
12   storing data representing said first mass spectrum;
13 and
14   correlating said mass spectrum with a mass spectrum
15 in said library of spectra of known organisms to determine the
16 correspondence of said spectra with the spectra obtained from

41

17 peptides obtained from the organism to be identified, thereby
18 identifying said organism.

1   26. The method of claim 25, wherein the organism to
2   be identified is a bacterium, fungus or virus.

1   27. The method according to claim 25, wherein the
2   mixture of proteins is obtained by enzymatic digestion of the
3   organism's proteins.

1   28. A method for characterizing a carbohydrate
2   structure of interest from a mixture of carbohydrates,
3   comprising:
4   using a tandem mass spectrometer to receive a
5   plurality of carbohydrates obtained from the mixture of
6   carbohydrates, to select at least a first carbohydrate ion
7   from the mixture of carbohydrates in a first mass analyzer of
8   the tandem mass spectrometer, to fragment said first
9   carbohydrate and to generate a carbohydrate fragment mass
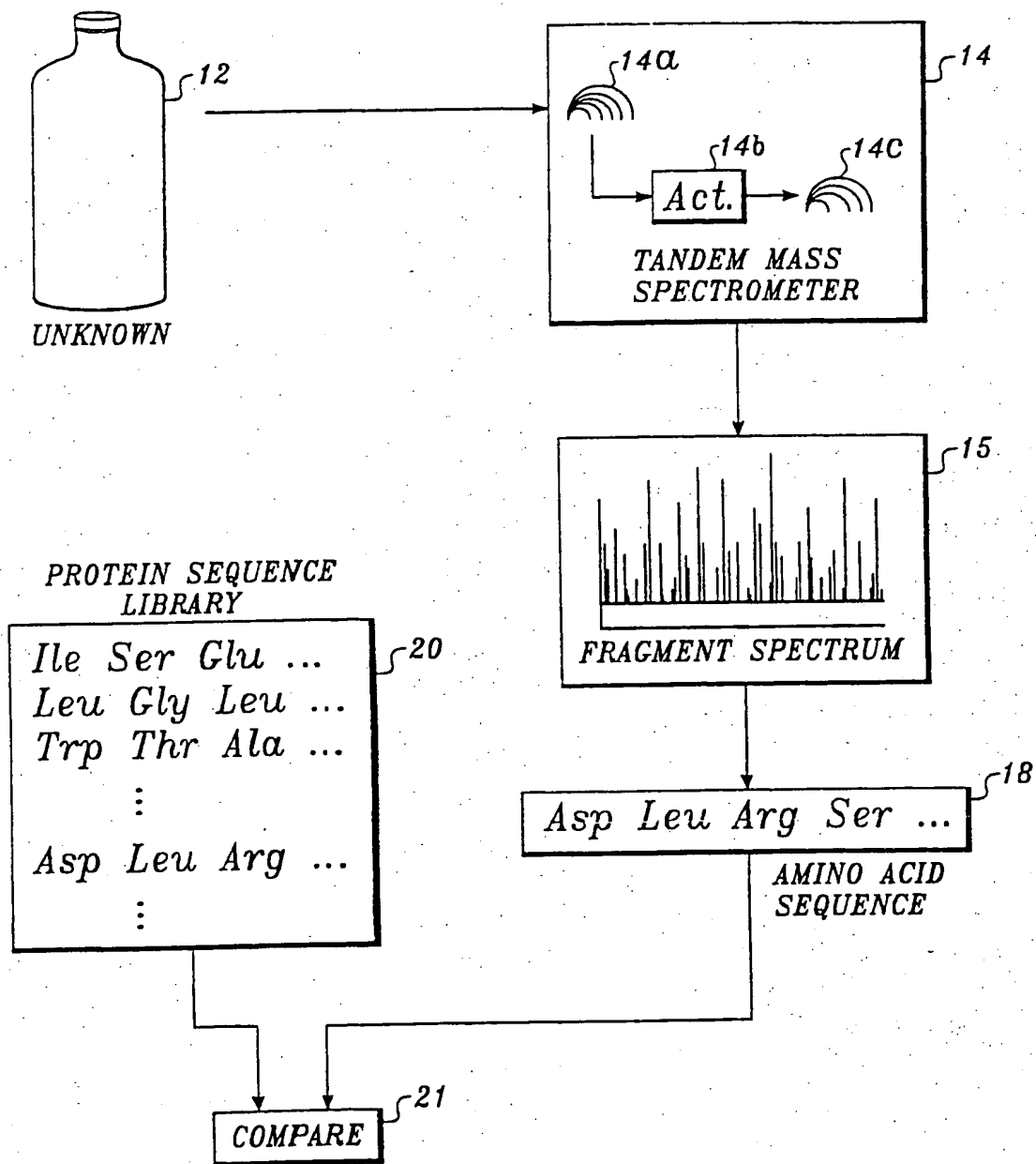10 spectrum;
11   storing data representing said first mass spectrum;
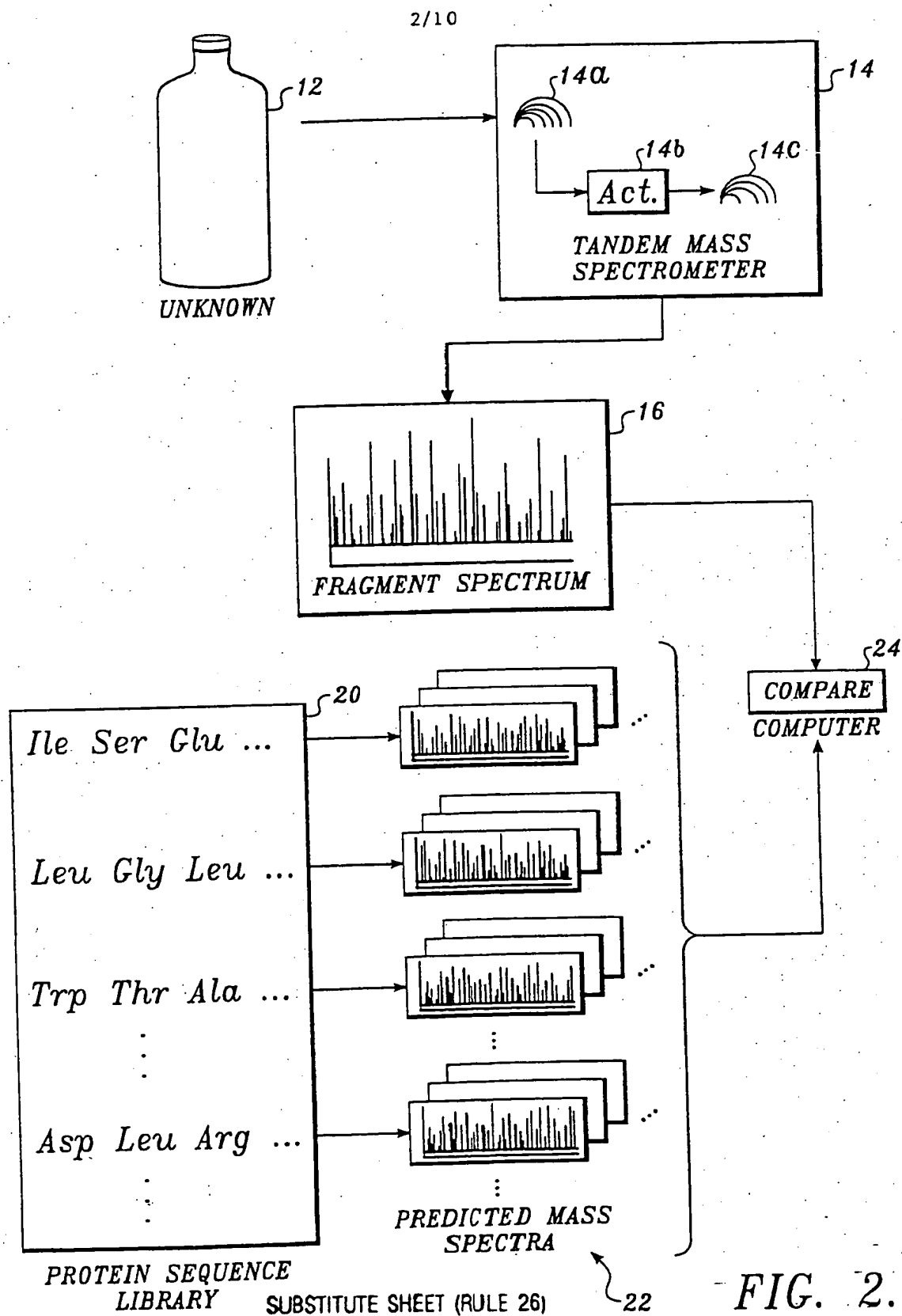12 and
13   correlating said mass spectrum with a database of
14 spectra of known carbohydrates, to determine the
15 correspondence of a carbohydrate specified in said
16 carbohydrate database with a carbohydrate in said mixture of
17 carbohydrates, thereby characterizing the structure of the
18 carbohydrate of interest.

1   29. The method of claim 28, wherein the mixture of
2   carbohydrates is obtained from a glycosylated protein of
3   interest.

1   30. The method of claim 29, wherein the mixture of
2   carbohydrates is obtained from a glycosylated protein of
3   interest by chemical or enzymatic release from the protein.

PROTEIN SEQUENCE
LIBRARY

UNKNOWN

TANDEM MASS
SPECTROMETER

FRAGMENT SPECTRUM

Asp Leu Arg Ser ...

AMINO ACID
SEQUENCE

Ile Ser Glu ...
Leu Gly Leu ...
Trp Thr Ala ...
⋮
Asp Leu Arg ...
⋮

COMPARE

(PRIOR ART)

FIG. 1.

UNKNOWN

14a

14

14b Act. *14c*

TANDEM MASS
SPECTROMETER

16

FRAGMENT SPECTRUM

24 COMPARE
COMPUTER

20

Ile Ser Glu ...

Leu Gly Leu ...

Trp Thr Ala ...

Asp Leu Arg ...

PROTEIN SEQUENCE
LIBRARY

PREDICTED MASS
SPECTRA

22

SUBSTITUTE SHEET (RULE 26)

FIG. 2.

FIG. 3.

4/10

$$INITIALIZE \ m \ = \ 0$$ —36

$$m \ = \ m \ + \ 1$$ —38

$$INITIALIZE \ SUM \ = \ 0$$
$$n \ = \ 0$$ —40

$$n \ = \ n \ + \ 1$$ —42

SUM = SUM + MW (m + n)
STORE SUM FOR USE
IN FRAGMENT CALCULATION —44

SUM <
[MW(TARGET) − TOLERANCE]
? —46

YES   NO

SUM <
[MW(TARGET) + TOLERANCE]
? —48

YES   NO

MARK THE m THROUGH
m + n
SUB−SEQUENCE AS A CANDIDATE —50

FIG. 4.

FIG. 5.

FIG. 6A
PREPROCESSING

ACQUIRE DATA FROM MASS SPECTROMETER — 602

SAVE DATA TO FILE AND CONVERT TO ASCII FORMAT — 604

START — 606

GET PEPTIDE MASS AND PRECURSOR ION CHARGE STATE FROM USER — 608

LOAD ASCII MASS/INTENSITY VALUES, ROUNDING TO UNIT MASSES — 610

REMOVE PRECURSOR ION CONTRIBUTION FROM DATA — 612

NORMALIZE REMAINING DATA TO MAXIMUM INTENSITY OF 100 — 614

NOTE PRESENCE OF IMMONIUM IONS H, F, AND Y — 616

STORE PEPTIDE MASS AND IMMONIUM ION INFO. IN DATAFILE — 618

SELECT TOP 200 MOST INTENSE PEAKS — 620

IF TWO PEAKS ARE WITHIN 2 AMU OF EACH OTHER, SET LOWER INTENSITY EQUAL TO THE GREATER INTENSITY — 622

STORE IN DATAFILE FOR PRELIMINARY SCORING — 624

DIVIDE DATA INTO 10 WINDOWS — 626

NORMALIZE TO MAXIMUM INTENSITY OF 50 WITHIN EACH WINDOW — 628

STORE IN DATAFILE FOR FINAL CORRELATION SCORING — 630

END — 632

SUBSTITUTE SHEET (RULE 26)
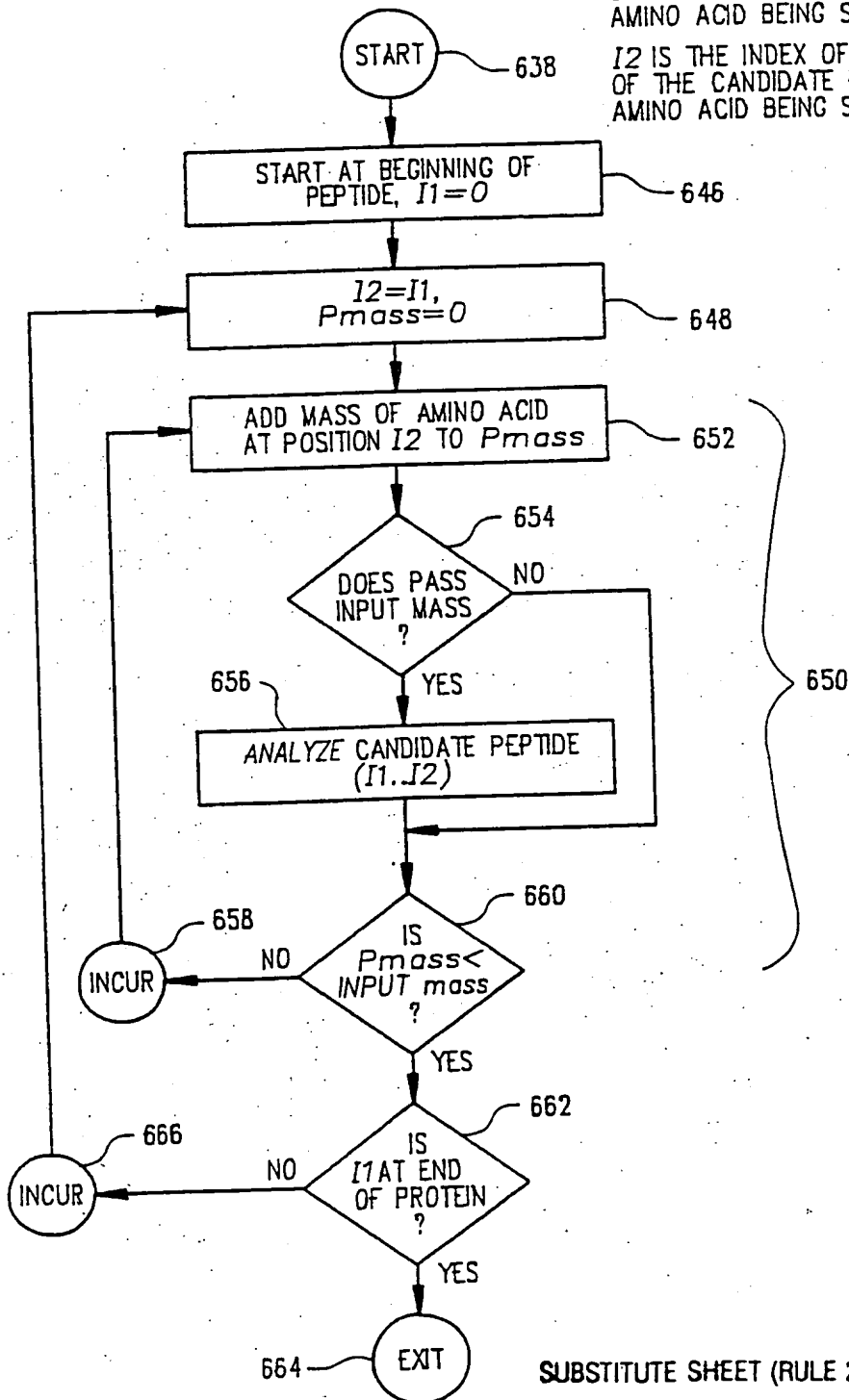
FIG. 6B
DATABASE SEARCH

FIG. 6C
SEARCH

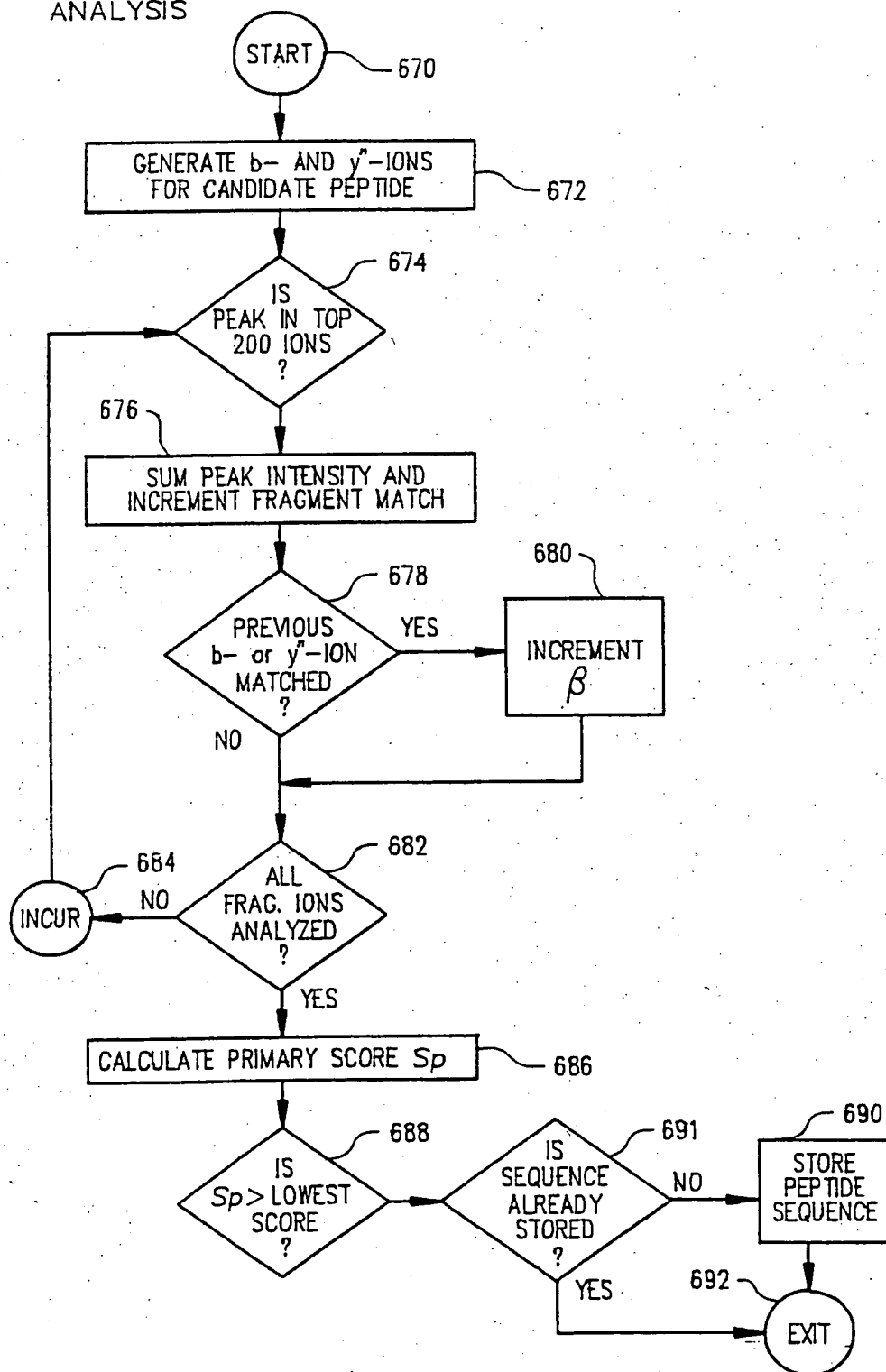*Pmass* IS THE MASS OF THE CANDIDATE PEPTIDE AS IT IS BEING SUMMED.

I1 IS THE INDEX OF THE START POSITION OF THE CANDIDATE PEPTIDE WITHIN THE AMINO ACID BEING SEARCHED.

I2 IS THE INDEX OF THE END POSITION OF THE CANDIDATE PEPTIDE WITHIN THE AMINO ACID BEING SEARCHED.
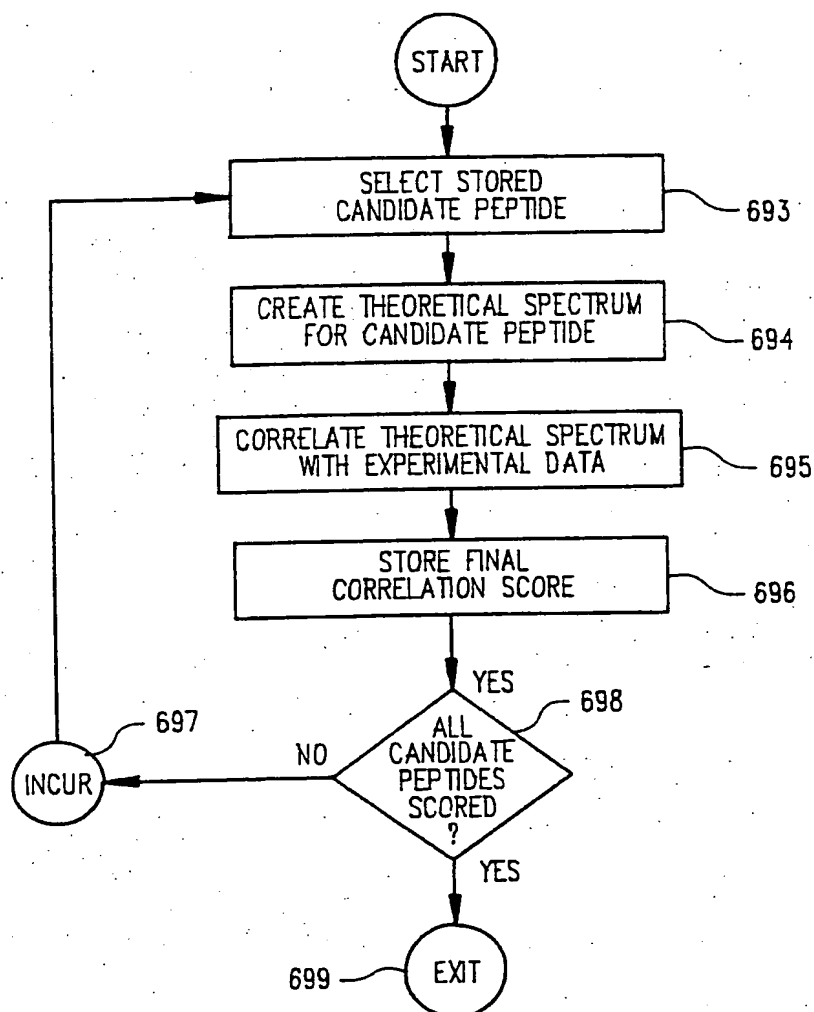
START — 638

START AT BEGINNING OF PEPTIDE, $I1=0$ — 646

$I2=I1$, $Pmass=0$ — 648

ADD MASS OF AMINO ACID AT POSITION $I2$ TO $Pmass$ — 652

654

DOES PASS INPUT MASS ? — NO

YES

656 — *ANALYZE* CANDIDATE PEPTIDE $(I1..I2)$

650

658 — INCUR — NO

660

IS $Pmass <$ INPUT mass ?

YES

666 — INCUR — NO

662

IS $I1$ AT END OF PROTEIN ?

YES

664 — EXIT

FIG. 6D
ANALYSIS

FIG. 6E

CORRELATION ANALYSIS

INTERNATIONAL SEARCH REPORT

A.   CLASSIFICATION OF SUBJECT MATTER

IPC(6)   :G01N 33/00

US CL   :436/89, 94

According to International Patent Classification (IPC) or to both national classification and IPC

B.   FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. :   Please See Extra Sheet.

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C.   DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PROC. NATL. ACAD. SCI. USA, VOLUME 90, ISSUED JUNE 1993, HENZEL, ET AL., "IDENTIFYING PROTEINS FROM TWO-DIMENSIONAL GELS BY MOLECULAR MASS SEARCHING OF PEPTIDE FRAGMENTS IN PROTEIN SEQUENCE DATABASES," PAGES 5011-5015.	1-17,20,21
A	PROC. NATL. ACAD. SCI. USA, VOLUME 83, ISSUED SEPTEMBER 1986, HUNT, ET AL., "PROTEIN SEQUENCING BY TANDEM MASS SPECTROMETRY", PAGES 6233-6237, SEE PAGES 6236-6237.	1-17,20,21
A	ANALYTICAL BIOCHEMISTRY, VOLUME 214, ISSUED 1993, YATES III, ET AL., "PEPTIDE MASS MAPS: A HIGHLY INFORMATIVE APPROACH TO PROTEIN IDENTIFICATION", PAGES 1-12, SEE ENTIRE ARTICLE.	1-17, 20, 21

[X] Further documents are listed in the continuation of Box C.   [ ] See patent family annex.

*   Special categories of cited documents:

"A"   document defining the general state of the art which is not considered to be of particular relevance

"E"   earlier document published on or after the international filing date

"L"   document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O"   document referring to an oral disclosure, use, exhibition or other means

"P"   document published prior to the international filing date but later than the priority date claimed

"T"   later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X"   document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y"   document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"   document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 JUNE 1995	12 JUL 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DAVID REDDING Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ANALYSIS OF PROTEINS BY MASS SPECTROMETRY, ISSUED 1992, GRIFFIN, ET AL., "ANALYSIS OF PROTEINS BY MASS SPECTROMETRY", PAGES 467-476, SEE PAGES 471-476.	1-17, 20, 21
A	J. AM. SOC. MASS SPECTROM., VOLUME 3, ISSUED 1992, McLUCKEY, ET AL., "TANDEM MASS SPECTROMETRY OF SMALL, MULTIPLY CHARGED OLIGONUCLEOTIDES", PAGES 60-70, SEE ENTIRE ARTICLE.	18-19

INTERNATIONAL SEARCH REPORT

B. FIELDS SEARCHED
Minimum documentation searched
Classification System: U.S.

436/89, 94, 173
435/6, 89, 91
530/334-337, 402, 412, 417